Bioinformatics, 37(5), 2021, 650–658
doi: 10.1093/bioinformatics/btaa852
Advance Access Publication Date: 5 October 2020
Original Paper

Gene expression

IntAPT: integrated assembly of phenotype-specific transcripts from multiple RNA-seq profiles

Xu Shi [1,2,*], Andrew F. Neuwald[3], Xiao Wang[1], Tian-Li Wang[4], Leena Hilakivi-Clarke[5], Robert Clarke[5] and Jianhua Xuan [1,*]

[1]Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA 22203, USA, [2]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA, [3]Institute for Genome Sciences and Department of Biochemistry & Molecular Biology, University of Maryland School of Medicine, Baltimore, MD 21201, USA, [4]Department of Pathology, Johns Hopkins Medical Institutions, Baltimore, MD 21231, USA and [5]Hormel Institute, University of Minnesota, 801 16th Ave NE, Austin, MN 55912, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia
Received on May 19, 2019; revised on August 27, 2020; editorial decision on September 20, 2020; accepted on September 21, 2020

Abstract

Motivation: High-throughput RNA sequencing has revolutionized the scope and depth of transcriptome analysis. Accurate reconstruction of a phenotype-specific transcriptome is challenging due to the noise and variability of RNA-seq data. This requires computational identification of transcripts from multiple samples of the same phenotype, given the underlying consensus transcript structure.

Results: We present a Bayesian method, integrated assembly of phenotype-specific transcripts (IntAPT), that identifies phenotype-specific isoforms from multiple RNA-seq profiles. IntAPT features a novel two-layer Bayesian model to capture the presence of isoforms at the group layer and to quantify the abundance of isoforms at the sample layer. A spike-and-slab prior is used to model the isoform expression and to enforce the sparsity of expressed isoforms. Dependencies between the existence of isoforms and their expression are modeled explicitly to facilitate parameter estimation. Model parameters are estimated iteratively using Gibbs sampling to infer the joint posterior distribution, from which the presence and abundance of isoforms can reliably be determined. Studies using both simulations and real datasets show that IntAPT consistently outperforms existing methods for the IntAPT. Experimental results demonstrate that, despite sequencing errors, IntAPT exhibits a robust performance among multiple samples, resulting in notably improved identification of expressed isoforms of low abundance.

Availability and implementation: The IntAPT package is available at http://github.com/henryxushi/IntAPT.

Contact: xuan@vt.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Alternative splicing is an important biological process transforming pre-mRNA into variant mRNA transcripts (Shi, 2017), more than 90% of which undergo alternative splicing in humans (Wang *et al.*, 2008). Reconstructing these transcripts, collectively termed the transcriptome, is crucial for understanding complex biological systems (Martin and Wang, 2011). High-throughput RNA sequencing (RNA-seq) technologies have revolutionized transcriptome analysis. By deep sampling of the transcriptome, RNA-seq can both quantify each transcript and identify isoforms, alternative splice junctions and gene fusions (Conesa *et al.*, 2016). However, inference of RNA transcripts from short-read technologies is difficult for many reasons, one of which is the number of candidate isoforms due to alternative splicing. Even more challenging is the reconstruction of isoforms with low coverage or with subtle structural differences due to the high variability of RNA-seq read distribution and to noise (McIntyre *et al.*, 2011).

Two categories of transcriptome assembly algorithms, *de novo* and reference-based (*ab initio*), currently address this problem. *De novo* algorithms directly assemble the transcriptome from raw RNA-seq reads (Holzer and Marz, 2019). Unlike *de novo* assemblers, reference-based assemblers take advantage of the existing high-quality reference genome by first aligning the short reads to the reference genome using RNA aligners, such as TopHat2 (Kim *et al.*, 2013) and STAR (Dobin *et al.*, 2013). Compared with the *de novo* strategy, reference-based assemblers lower the complexity of the problem by first grouping the reads based on matching genomic locations (Martin and Wang, 2011). Ideally, each group would represent the reads coming from a single gene. RNA-seq aligners

identify splice junctions between exons in each gene based on splice site-mapped reads. Reference-based assemblers construct splicing graphs using exons as nodes and splice junctions as edges. Paths in the splicing graph represent potential transcripts. The key challenge is to first determine the relationship between observed reads and (unobserved) transcripts and to then infer the existence and abundance of each transcript. Current reference-based assemblers, including Cufflinks (Trapnell *et al.*, 2010), IsoLasso (Hah *et al.*, 2011), CEM (Li and Jiang, 2012), SLIDE (Li *et al.*, 2011) and SparseIso (Shi *et al.*, 2018), apply sparsity-enforced methods to provide biologically plausible solution to the problem, which alleviates overfitting due to the high complexity of the graph.

Although existing algorithms can effectively identify transcript isoforms from a single RNA-seq sample, due to the variation of splicing profiles between samples and the sampling bias of RNA-seq technology (McIntyre *et al.*, 2011), analysis of a single sample may not capture the transcriptome comprehensively. Several research projects, including ENCODE (Djebali *et al.*, 2012) and TCGA (Cancer Genome Atlas Network, 2012), have produced a large amount of RNA-seq data. The sizes of these datasets now make it possible to study the phenotype-specific transcriptome assembly using multiple RNA-seq profiles. A simple way to identify transcripts from multiple samples is to apply single-sample isoform identification methods to merged or pooled reads from multiple samples. However, noisy reads accumulate during merging, thereby hindering the identification of low abundance isoforms or junctions. An alternative approach is to identify single-sample isoforms prior to merging the samples. Within the Cufflinks package, the Cuffmerge program uses a similar idea to assemble transcripts. Cuffmerge first uses Cufflinks to assemble transcripts from each sample, and then, treating the assembled transcripts as long reads, generates an artificial dataset. Finally, Cuffmerge uses Cufflinks again to process the artificial data for transcript identification. As the number of samples increases, erroneous isoforms identified from individual samples accumulate, thereby decreasing the accuracy of both isoform identification and quantification (Niknafs *et al.*, 2017; Tasnim *et al.*, 2015).

Recently, TACO (Niknafs *et al.*, 2017) has been developed to address Cuffmerge's high false positive rate. TACO uses a change-point detection strategy to accurately identify the transcript start and end sites, which leads to improved assembly accuracy and to significantly fewer false positives. However, this approach can filter out many isoforms of low abundance. Several other methods, such as Iterative Shortest Path (ISP) (Tasnim *et al.*, 2015) and FlipFlop (Bernard *et al.*, 2015), have been developed for transcriptome assembly from multiple RNA-seq samples. Unlike Cuffmerge, these methods directly infer the isoform structure from multiple samples. These methods first construct a multiple sample splicing graph (MSSG) from the data with exons as nodes and junctions as edges. Nodes and edges are then assigned weights based on their depth of coverage. Finally, sparsity enforcing methods find the paths or isoforms on the graph with the highest weight. However, these approaches tend to miss many isoforms of low abundance.

Here, we use a Bayesian approach, integrated assembly of phenotype-specific transcripts (IntAPT), to infer phenotype-specific isoforms from multiple RNA-seq profiles by developing a hierarchical version of SparseIso. The key idea is to correctly predict phenotype-specific isoforms given a consensus transcriptome for the phenotype with the underlying consensus transcriptome profiles helping to support transcript identification. The method experiment would be more focused on the transcriptome profiles of this phenotype in general rather than on each individual cell, which should reduce noise and improve reproducibility. Bayesian inference is used to predict those isoforms and their abundances most likely to have generated the observed multiple-sample RNA-seq data; by avoiding expression thresholds, this process allows detection of lowly expressed isoforms. More specifically, we first build an MSSG using the ISP algorithm and enumerate candidate isoforms corresponding to maximal paths through the graph. We next use a two-layer Bayesian model to infer isoforms from the observed reads. This captures the dependency of the abundance of isoforms (sample level) on the presence of isoforms (group level). At the group level, we categorize candidate isoforms into 'unexpressed' and 'expressed' groups. Isoforms in the unexpressed group have expression levels near zero. Reads from the unexpressed group are likely due to noise generated by various errors, including sequencing or mapping errors. Each isoform's group state is modeled as a Bernoulli random variable with a high prior probability for the unexpressed state, which enforces the sparsity of isoforms. The sparsity constraint helps alleviate overfitting of the observed reads. Model parameters are estimated from the joint posterior distribution using Gibbs sampling. We iteratively estimate the presence of each isoform at the group level and the corresponding abundance at the sample level. The final sets will only include isoforms estimated with high confidence. To evaluate performance, we compared IntAPT with existing methods using both simulated data and real data. The results show that IntAPT consistently outperforms popular methods for phenotype-specific transcriptome assembly.

2 Materials and methods

The flowchart of the IntAPT method is shown in Figure 1. We first align sequence reads to the reference genome. Next, we identify exons and introns using the coverage of reads along the genome and the junctions between exons using spliced reads that map to multiple exons. We then combine the identified exons and junctions from all the samples to construct a MSSG. Candidate transcript isoforms are enumerated as all maximal paths through the graph. We model the expression values for each exon and junction in the graph using a two-parameter negative binomial (NB) distribution. This NB model addresses the over-dispersion problem observed for RNA-seq data (Robinson *et al.*, 2010; Robinson and Smyth, 2007). Expression of exons and junctions is modeled in this way as mixtures of isoform expressions.

We use a two-layer Bayesian framework (group layer and sample layer) to model the observed reads from multiple RNA-seq samples. At the group layer, we use a binary variable to indicate the presence of each candidate isoform. Isoform expression in the sample layer depends on the group-level presence. If an isoform is labeled as unexpressed, its expression value will approach zero. Otherwise, the expression value will be >0. The two-layer structure of the parameters in the model helps estimate phenotype-specific isoforms and their level of expression. When we evaluate a given isoform in one sample, evidence for this isoform in other samples will increase the probability of its occurrence. Because sequencing and mapping errors are less likely to be replicated in all RNA-seq samples, this also helps eliminate false positives. We use Gibbs sampling to estimate the parameters of the posterior distribution; when the isoform state is more accurately estimated in this way, weakly expressed isoforms are more likely to be detected. The confidence level of each candidate isoform is estimated by the frequency with which it is sampled as being expressed. Therefore, by providing both the abundance of each candidate isoform and a corresponding confidence level, IntAPT allows researchers to prioritize the selection of isoforms as assembled.

2.1 Splicing graph construction

Based on the genomic location of read alignments, we use the processsam program in the CEM package (Kimmig *et al.*, 2012) to identify gene regions, by clustering single-sample reads, and splice junctions, as reported by RNA-seq aligners. Within each gene region, exons and introns are identified based on read coverage. We expect that the coverage of introns should be close to zero. Splice junctions between exons are identified from the spliced reads. Due to sequencing and mapping errors, the identified splice junctions in each sample are filtered by the number of supporting reads. We build a candidate set of phenotype-specific splice junctions as the union of identified splice junctions from all samples, from which is obtained the total number of supporting reads. Exons and junctions with poorly supported samples are filtered out, since these likely arise from sequencing or mapping errors.

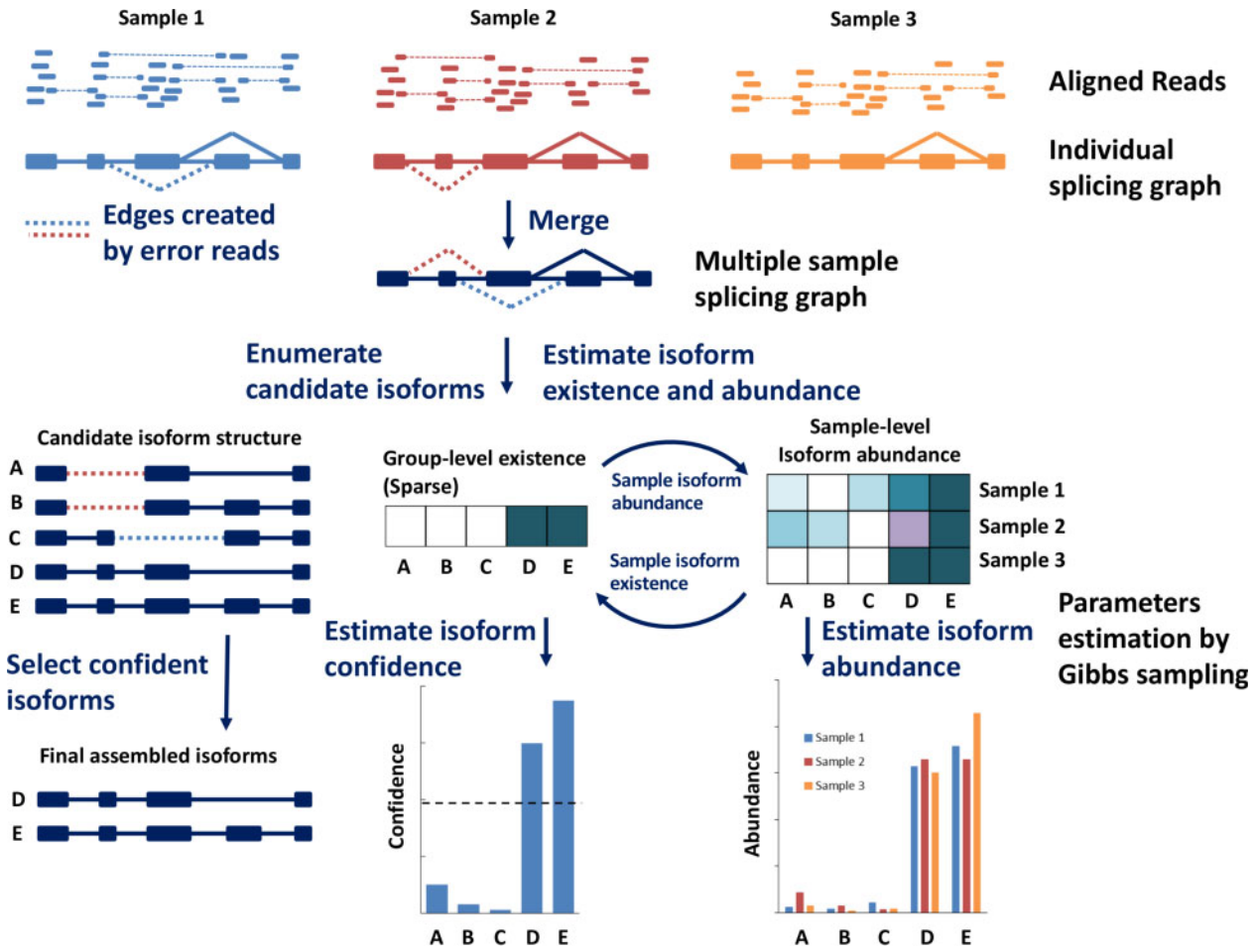

Fig. 1. Flowchart of the IntAPT method for phenotype-specific isoform identification. (i) A MSSG is constructed from the splice junctions from all samples. The candidate isoforms are enumerated as the maximal paths of the MSSG. (ii) A two-layer Bayesian model is built to explain the observed reads as isoform group-level expression states and corresponding sample-level expression values. (iii) Gibbs sampling estimates the model parameters iteratively. Due to interdependencies, the expression of isoforms in different samples will support each other through the group layer. The final set of assembled isoforms is selected based on the estimated expression state

The MSSG, $G = (V, E)$, is built by connecting exons from the identified phenotype-specific junctions. The node set V also includes junctions. Junctions carry direct information about splicing, which helps to assign reads to candidate transcripts (see Supplementary Section S1.1). All candidate transcripts will be enumerated from the graph. For a detailed description of splicing graph construction see Supplementary Section S1.2.

2.2 Read count model

We use a two-parameter NB model to address the problem of over-dispersion. As described in the step for candidate isoform construction, we include junctions in our model to enhance interpretation of isoforms. We define segments as the union of exons and junctions. We assume that an observed read count $R_{i,m}$ (from segment i and sample m) was drawn from a NB distribution as follows:

$$R_{i,m} \sim \mathrm{NB}\left(E_{i,m} l_i, \tau_i^2\right), \tag{1}$$

with a mean of $E_{i,m}$ and variance of $E_{i,m}(1+\tau_i^2)$. $E_{i,m}$ is the expression of segment i in sample m. τ_i^2 is a parameter controlling the over-dispersion of read counts on segment i. When $\tau_i^2 = 0$, the NB distribution is equivalent to a Poisson distribution. l_i is the effective length of segment i and represents the expected number of bases that support reads mapped entirely within the segment (see Supplementary Section S1.3 for details).

2.3 Hierarchical Bayesian model for transcript inference from multiple samples

Just as the read counts observed at an exon can be modeled as a mixture of counts generated from multiple transcripts, the segments expressed across multiple samples can be modeled as a linear mixture of expressed transcripts:

$$\boldsymbol{E} = \boldsymbol{X}\boldsymbol{e} + \boldsymbol{\sigma}^2 \sim \prod_m N_+\left(E_m | \boldsymbol{X}\boldsymbol{e}_m, \sigma_m^2 \boldsymbol{I}\right), \tag{2}$$

where $\boldsymbol{X}$ is a matrix of binary indicator variables, such that $X(i,t) = 1$, when transcript t covers segment i, and $X(i,t) = 0$, otherwise. N_+ denotes the normal distribution truncated at 0. $\boldsymbol{X}\boldsymbol{e}_m$ and $\sigma_m^2 \boldsymbol{I}$ are the mean and variance of the normal distribution before truncation. $\boldsymbol{E}_m$ and $\boldsymbol{e}_m$ are the segment and transcript levels of expression, respectively, in sample m. $\boldsymbol{I}$ is an identity matrix. σ_m^2 is the sample-level variance following an inverse gamma distribution, which is the conjugate prior for the variance of a normal distribution.

Due to mapping uncertainty and transcriptome complexity, the number of candidate transcripts obtained from the MSSG is usually large. Therefore, we use a joint spike-and-slab prior (Ishwaran and Rao, 2005; Mitchell and Beauchamp, 1988) in conjunction with the two-layer Bayesian model to estimate the presence and expression of isoforms. This approach alleviates the problem of overfitting by enforcing sparsity. At the group layer, we introduce the binary variable $\boldsymbol{w} = [w_1, w_2, \ldots, w_T]$ to indicate whether or not each phenotype-specific isoform is expressed and $\boldsymbol{\gamma}^2 = [\gamma_t^2, \gamma_t^2, \ldots, \gamma_T^2]$ to indicate the variance of transcript expression across samples. If

transcript t is expressed, $w_t = 1$; otherwise $w_t = w_0$, where w_0 is a constant close to 0. w_t follows a Bernoulli distribution with parameter w:

$$w_t \sim \pi^{w_t}(1-\pi)^{w_t - w_0}, t = 1, 2, \ldots, T, \quad (3)$$

where w_t will have values of w_0 or 1 and T is the number of candidate isoforms. π is the prior probability for an isoform to express. Given the group-level variables, the prior distribution of the transcript expression across multiple samples are modeled jointly with spike-and-slab prior

$$\boldsymbol{e_t}|w_t \tilde{} N_+\left(\boldsymbol{e_t}|0, w_t\gamma_t^2\right) = \prod_m \frac{2}{\sqrt{2\pi w_t\gamma_t^2}} \exp\left(-\frac{e_{t,m}^2}{2w_t\gamma_t^2}\right) I(e_{t,m} > 0), \quad (4)$$

where $e_{t,m}$ is the abundance of transcript t in sample m, γ_t^2 models the variability of isoform abundance at the group level and $I(x)$ is an indicator function. The truncated normal distribution of $\boldsymbol{e}$ guarantees conjugacy, as a normal distribution is the conjugate prior distribution of a normal distribution with known variance. Transcript expressions in multiple samples are controlled by the same group-level parameters w_t and γ_t^2. For an unexpressed transcript ($w_t = w_0$), the corresponding variance of isoform expression will be small, resulting in an expression value near zero for most samples. For expressed transcripts ($w_t = 1$), expression values can be larger than zero. If isoform t is expressed in some samples, w_t and γ_t^2 are likely to be large, which will support the estimated abundance of isoform t across all samples. For the conjugacy of the Bayesian model, we let γ_t^2 follow an inverse Gamma distribution. Examples of the spike-and-slab prior can be found in Supplementary Section S1.4. To further understand the two-layer model, the marginal prior of $\boldsymbol{e_t}$ can be derived by integrating out nuisance variables:

$$\boldsymbol{e_t} \sim \left[\pi * t_{2a}\left(\boldsymbol{e_t}|0, \frac{b}{a}\boldsymbol{I}\right)\right]^{w_t} \left[(1-\pi) * t_{2a}\left(\boldsymbol{e_t}|0, \frac{w_0 b}{a}\boldsymbol{I}\right)\right]^{w_t - w_0} I(\boldsymbol{e_t} \geq 0), \quad (5)$$

where t_{2a} denotes the Student t distribution with degree of freedom $2a$, and a and b are the shape and rate of the prior distribution of γ_t^2 (inverse gamma distribution). Transcript expressions in individual samples are independent only if $w_0 = 1$, which is not the case here. Therefore, expression variation dependencies among transcripts are implicitly modeled through the two-layer Bayesian model. As described in Supplementary Section S1.5, there are positive correlations among individual transcript expressions, which govern the consistency of the phenotype-specific expression pattern. Specifically, we show that the joint distribution of isoform expressions follows a mixture of Student's t-distributions, which can be mathematically derived based on an important property of the two-layer Bayesian model: given the group-level presence and variance at the group level, isoform expressions at the sample level are conditionally independent.

2.4 Model parameter estimation using Gibbs sampling

We use Gibbs sampling to estimate the parameters of the Bayesian model given segment expression $\boldsymbol{E}$. The relationship between the parameters of the model is shown in Figure 2. By iteratively drawing samples from conditional distributions, we can infer the joint posterior distribution of the model variables and parameters. The final outputs of the Gibbs sampler are the confidence level of each isoform's presence and its corresponding abundance. The full joint posterior distribution can be described as follows:

$$P(\boldsymbol{e}, \boldsymbol{\gamma}^2, \boldsymbol{w}, \boldsymbol{\sigma}^2, \pi|\boldsymbol{E}, \boldsymbol{X}) \sim P(\boldsymbol{E}|\boldsymbol{X}, \boldsymbol{e}, \boldsymbol{\sigma}^2) \times P(\boldsymbol{e}|\boldsymbol{\gamma}^2, \boldsymbol{w}) \times P(\boldsymbol{\gamma}^2) \times P(\sigma^2) \times P(\boldsymbol{w}) \times P(\pi). \quad (6)$$

In the sample layer, given the group layer parameters, M samples and T candidate transcripts, we perform sampling of the isoform expression in sample m independently from

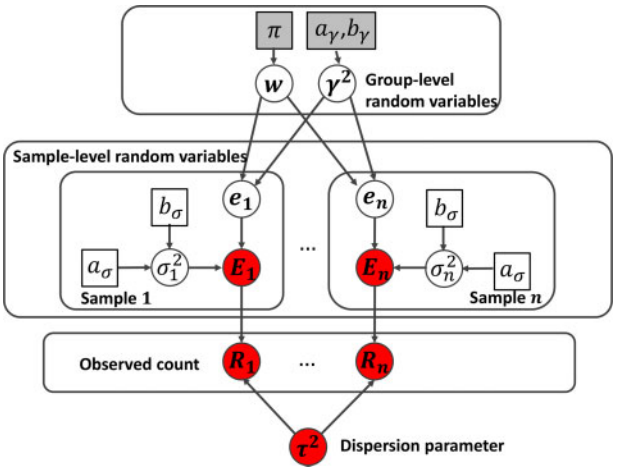

Fig. 2. Dependency graph of the IntAPT model. The known and unknown variables are colored red and white, respectively. Variables in squares are hyperparameters

$$P(\boldsymbol{e_m}|\boldsymbol{E_m}, \boldsymbol{X}, \boldsymbol{\gamma}^2, \boldsymbol{w}, \sigma_m^2) \sim N_+\left(\boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{E_m}, \sigma_m^2\boldsymbol{A}^{-1}\right), \quad (7)$$

where $\boldsymbol{E_m}$ is the segment expression, $\boldsymbol{A} = \boldsymbol{X}^T\boldsymbol{X} + \sigma_m^2\boldsymbol{D}_\gamma^{-1}$, and $\boldsymbol{D}_\gamma$ is a diagonal matrix whose diagonal elements are $[w_1\gamma_1^2, \ldots, w_T\gamma_T^2]$. Here, $\boldsymbol{X}^T\boldsymbol{X}$ encodes the isoforms structural consistency, where more shared segments among isoforms lower the probability that they are expressed simultaneously. The group-level expression state $\boldsymbol{w}$ has an effect on both mean and variance. If isoform t is expressed, its abundance tends to be high with the support of w_t. The truncated normal distribution is simulated by the Gibbs sampler introduced in Damien and Walker (2001) (see Supplementary Section S1.6 for more details). Similar to the prior distribution, the conditional distribution of $\boldsymbol{w}$ is also a mixture of w_0 and 1. The elements of $\boldsymbol{w}$ are sampled independently from

$$P\left(w_t|\gamma_t^2, \pi, \boldsymbol{e_t}\right) \tilde{} p_1^{w_t}(1-p_1)^{w_t - w_0}, \quad (8)$$

where $\boldsymbol{e_t}$ is the expression of isoform t across all samples, $p_1 = k_1/(k_1 + k_2)$, $k_1 = \pi\exp(-\sum_{m=1}^M e_{t,m}^2/2\gamma_t^2)$ and $k_2 = (1-\pi)(w_0)^{-M/2}\exp(-\sum_{m=1}^M e_{t,m}^2/2w_0\gamma_t^2)$ and $e_{t,m}$ is the expression of isoform t in sample m. Weights of the spike and slab of w_t are determined by fitting the estimated expression of isoform t to the spike and slab in the group level. The iterative framework described above further illustrates the connection between the sample layer and group layer in our Bayesian model, which improves the estimation of both the presence and abundance of isoforms. The noise $\boldsymbol{\sigma}^2$ in the sample layer and the variance $\boldsymbol{\gamma}^2$ can be sampled from an inverse Gamma distribution, with parameters derived from the corresponding conditional distribution. In the sampling framework, we estimate π from its conditional distribution. Derivation of the conditional distributions and a detail sampling procedure are given in Supplementary Section S1.7.

2.5 Implementation and availability

The IntAPT algorithm is implemented as a C++ package, which is made available to the research community at http://github.com/henryxushi/IntAPT. Note that, the improved performance of IntAPT was achieved without sacrificing computational time (see Supplementary Section S1.8).

3 Results

3.1 Benchmarking for performance evaluation

We compared the performance of IntAPT with that of the most widely used tools available: Cufflinks-pool (Cufflinks v2.1.1),

Cuffmerge (in the package of Cufflinks v2.1.1), ISP (v0.3), FlipFlop (v1.10), TACO (v0.6.3) and StringTie (v1.3.3b) under merge mode. Cufflinks-pool is a simple, alternative way to identify isoforms from multiple samples using the Cufflinks package. To run Cufflinks-pool, we first merged the reads from all samples and then applied Cufflinks to the pooled data (treated as a single RNA-seq sample) for transcript assembly. All assemblers were run with default parameters, except for the mean and SD of fragment length distribution in FlipFlop, which was set according to the results from Cufflinks. We used Cufflinks assembled transcripts as the input for TACO. Because the parameters of IntAPT are estimated using Gibbs sampling, the results of different runs will undergo slight differences; therefore, we used the median performance of five independent runs.

Sensitivity and precision are widely used benchmarks to evaluate the performance of transcript assemblers. Sensitivity is defined as the fraction of successfully assembled isoforms in the reference set; precision is defined as the fraction of predicted isoforms that are present in the reference set. Besides sensitivity and precision, we also compute an *F*-score, the harmonic mean of precision and sensitivity [calculated as 2×precision×sensitivity/(precision+sensitivity)], to evaluate the overall performance of assemblers. An identified isoform is considered to be assembled correctly if we can find an intron-chain match in the reference set. We used the Cuffcompare tool in the Cufflinks package (Trapnell *et al.*, 2010) (which uses the same matching strategy) to find the intron-chain matching between the predicted isoforms and the reference set.

3.2 Performance evaluation using simulation data

For simulated data, we first generated expression profiles of isoforms annotated in RefSeq (Pruitt *et al.*, 2014) [downloaded from the UCSC Genome Browser (Rosenbloom *et al.*, 2015)] using the Flux Simulator (Griebel *et al.*, 2012). Expression values of simulated isoforms across multiple samples follow a Gamma distribution. Sequence reads were then generated by RNASeqReadSimulator (http://alumni.cs.ucr.edu/~liw/rnaseqreadsimulator.html) as used in Djebali *et al.* (2012). For each sample, we generated 50 million 100 bp paired-end reads of simulated RefSeq isoforms from the GRCh37/hg19 human reference genome. In total, we simulated six samples in our dataset. Due to technical and sampling variability, the same set of isoforms might not be expressed in all samples (McIntyre *et al.*, 2011). For each simulated sample, we randomly assigned ~11 000 transcripts as truly expressed, which is comparable to the number of transcripts identified in previous studies (Pertea *et al.*, 2015). In the expression profile from the Flux Simulator, genes had different numbers of expressed isoforms (labeled as N_e), as the isoforms were randomly selected from RefSeq. The distribution of N_e is shown in Supplementary Figure S2-1. Due to the complexity of the splicing graph, isoforms of genes with a higher N_e are more difficult to assemble. For example, the splicing graph of one gene will be a single chain if $N_e = 1$. As N_e increases, more edges will be added to make the graph more complicated. Due to the large difference in splicing graph complexity, we simulated the genes with $N_e > 1$ and $N_e = 1$ separately. Importantly, in this study, our main focus is to assemble the transcripts of genes with $N_e > 1$, as they are among the most challenging to assemble. The results of genes with $N_e = 1$ are described in Supplementary Section S1.9.

Figure 3 shows the *F*-score, precision, and sensitivity of all assemblers on genes with $N_e > 1$. Based on *F*-scores, IntAPT exhibited a significant improvement over all competing methods. Compared with the next two algorithms, StringTie and FlipFlop, the performance of IntAPT had an increase in *F*-score of 4.82% (0.8248 versus 0.7869) and 9.68% (0.8248 versus 0.7520), respectively. Specifically, IntAPT correctly assembled about 15.10% and 7.37% more transcripts than Cuffmerge and StringTie, respectively. IntAPT's precision was similar to that of Cuffmerge, which was 11.88% (0.8062 versus 0.7206) higher than FlipFlop's. Due to a difference in strategy, Cufflinks-pool focuses more on highly expressed isoforms, which lowered its false positive rate. StringTie uses an iterative approach to search for maximum flow in the splicing graph, resulting in improved accuracy in predicting highly expressed isoforms. TACO's change-point detection strategy also lowered the false positive rate compared to Cuffmerge. However, the improved precision of Cufflinks-pool, StringTie and TACO also led to a relatively large portion of transcripts unidentified (low sensitivity), which affected the overall isoform identification performance as measured by *F*-score. Importantly, IntAPT correctly identified more transcripts than existing methods, independent of simulated expression level (Fig. 3D), as quantified in reads per kilobase of transcript per million mapped reads (RPKM). On highly expressed isoforms (RPKM >50), most assemblers performed well, achieving a sensitivity >0.8. Nevertheless, IntAPT and FlipFlop predicted about 10–30% more of lowly expressed isoforms (RPKM <10) than the other assemblers. Our Bayesian model was highly effective at assembling lowly expressed isoforms. Among existing methods, FlipFlop achieved the highest sensitivity, but this came with a relatively low precision, as it tended to predict large numbers of isoforms.

We also measured the Spearman's rank correlation between the simulated abundance and the predicted abundance of isoforms. Due to the different numbers of true isoforms identified by different algorithms, we used the isoforms correctly identified by all algorithms as the evaluation set. Based on this criterion, IntAPT performed better than existing methods in quantifying isoforms (Fig. 3E). To evaluate the performance comprehensively, we set different thresholds of abundance for each assembler to calculate a series of precisions and recalls (sensitivities). Supplementary Figure S2-2 shows the precision-recall curve and the corresponding area under the precision-recall curve (AUC). This reveals that IntAPT achieved an increase of 13.5% in terms of AUC compared with StringTie, a leading algorithm among the existing methods.

To demonstrate the effectiveness of transcriptome assembly methods using multiple RNA-seq profiles, we further evaluated the performance of all assemblers under different numbers of samples. Figure 4 and Supplementary Figure S2-3 show the *F*-score, sensitivity and precision. IntAPT had consistently higher *F*-scores and sensitivity when using different numbers of samples (ranging from 2 to 14). As the number of samples increased, the number of expressed isoforms also increased. Thus, the structure of the splicing graphs became more complex. With more samples, all assemblers identified more transcripts and achieved higher sensitivity, while the precision underwent a slight drop. Unlike the other methods, Cufflinks-pool directly works on pooled data and aims to identify a small set of isoforms. Therefore, Cufflinks-pool had high precision but with no improvement in sensitivity.

Sequencing errors and imperfect library preparation ensure that real data usually contain noisy reads. One screenshot of real cell line data is shown in Supplementary Figure S2-4 (Supplementary section S1.10), in which some noisy junction reads and intron reads are evident. To simulate RNA-seq data more realistically, we generated random error reads carrying false junctions and parts of introns. We evaluated the performance of all assemblers under different error rates, defined as the fraction of error reads (shown in Supplementary Fig. S2-5). As error reads increased, all assemblers had lower precision and sensitivity. However, the two-layer structure of IntAPT's Bayesian model tended to find consistent isoform sets across multiple samples, thereby decreasing the selection of erroneous isoforms; hence, IntAPT performed more robustly than existing methods (see Supplementary Fig. S2-5).

We further evaluated all assemblers on another realistic simulated dataset generated from a MCF7 cell line dataset from ENCODE (GEO accession number: GSM958745) using Polyester (Frazee *et al.*, 2015) (data generation is described in Supplementary Section S1.11.) ISP was not included in this analysis, as we found that it identified several thousands of erroneous isoforms that did not match any simulated structures. Supplementary Figure S2-6 shows *F*-scores, precision and recall. IntAPT achieved a 67.93% higher *F*-score than the second-best assembler, StringTie. Specifically, IntAPT achieved a substantial improvement in precision as the data generated by Polyester included more sequence

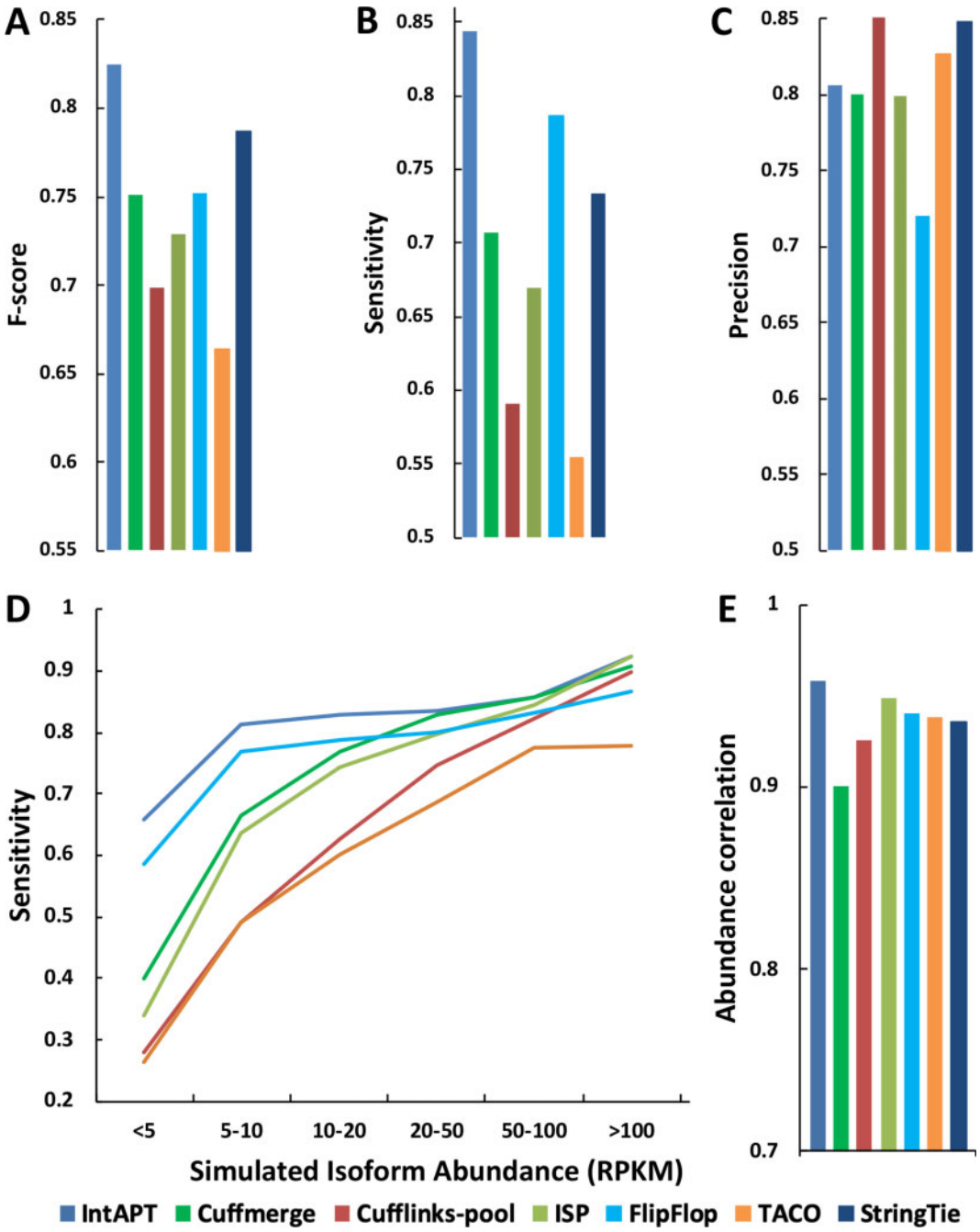

Fig. 3. Performance evaluation on simulated data of genes with $N_e > 1$. (**A**) Overall performance of isoform identification evaluated by *F*-score. (**B**) Sensitivity and (**C**) precision of all assemblers. (**D**) Sensitivity on isoforms with different levels of abundance quantified in FPKM. (**E**) Spearman's rank correlation between predicted abundance and simulated abundance

errors; this is consistent with the results of our other simulation studies with high error rates (Supplementary Fig. S2-5). IntAPT lowers the probability of selecting erroneous isoforms by considering the level of support between samples under the two-layer Bayesian framework.

3.3 Real RNA-seq data

To evaluate the ability of assemblers to identify phenotype-specific transcript isoforms using real RNA-seq data, we applied them to ENCODE datasets (Djebali *et al.*, 2012) from three different cell lines: human MCF-7 breast cancer cells, H1-hESC embryonic stem cells and HepG2 cells. Information on these cell line datasets is given in Supplementary Tables S2-1 and S2-2; preprocessing of the data is described in Supplementary Section S1.12, and the numbers of transcripts identified from these data are given in Supplementary Table S2-3.

Because the true set of expressed isoforms is neither available nor accessible for real RNA-seq data, there is currently no gold standard for directly evaluating isoform prediction in terms of precision and recall. However, we can use current annotation and other independent data sources to assess sensitivity and precision for known isoforms. We constructed a comprehensive reference isoform set by integrating known isoforms and Pacific Biosciences (PacBio) sequencing (Gonzalez-Garay, 2016; Rhoads and Au, 2015) data (see Supplementary Section S1.13 for details).

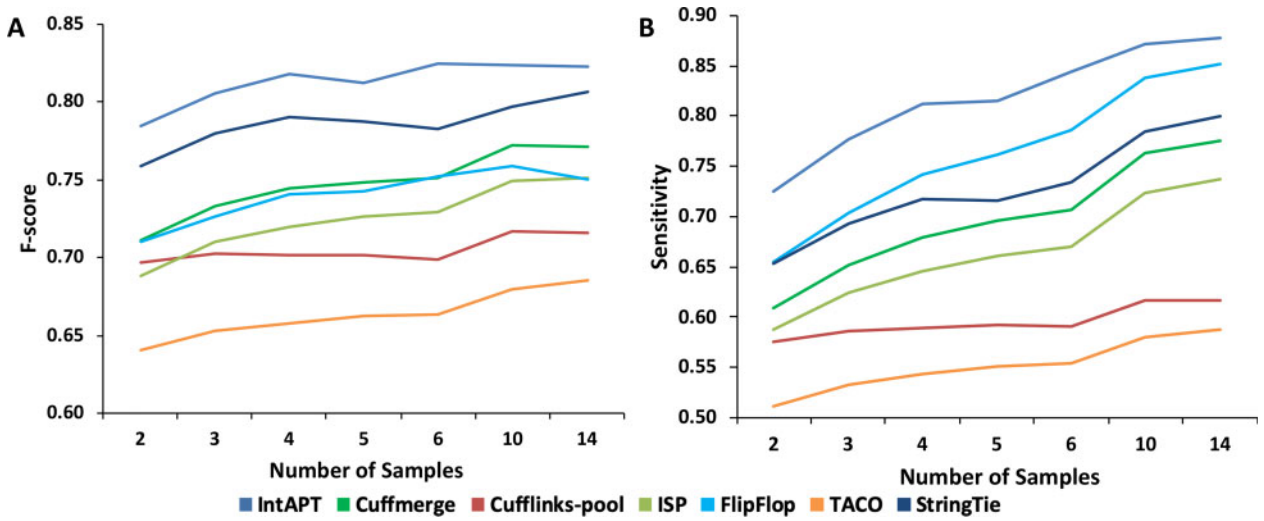

Fig. 4. Performance evaluation on simulation data of genes with N_e>1 under different number of samples in terms of (**A**) *F*-score and (**B**) sensitivity

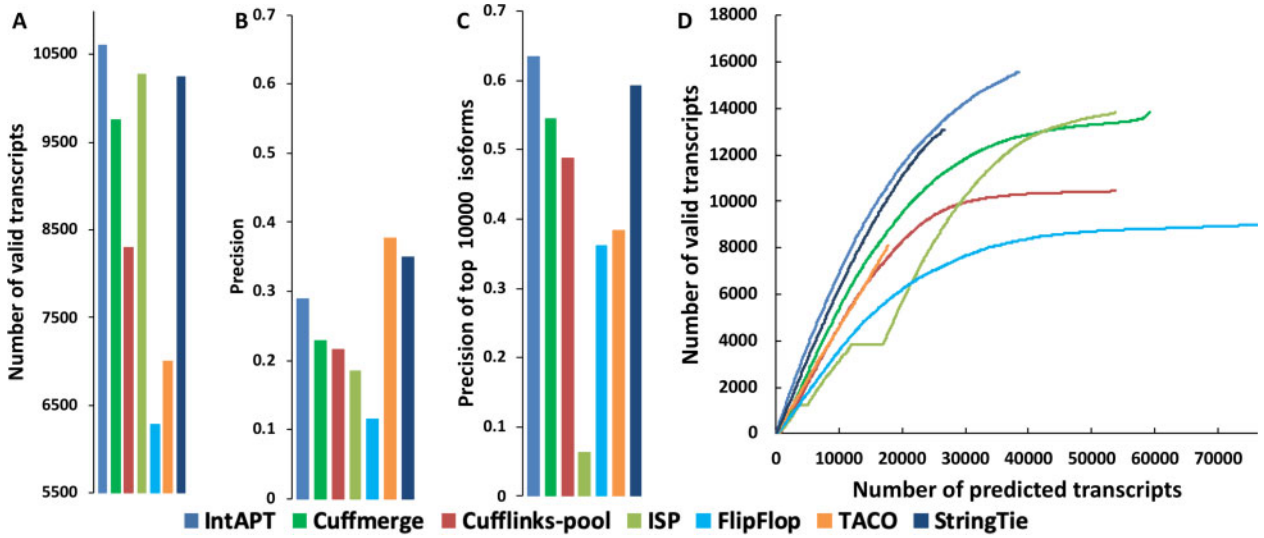

Fig. 5. Performance evaluation on MCF7 cell line data using PacBio transcriptome and RefSeq annotation: (**A**) number of valid transcripts, (**B**) precision, (**C**) precision of top 10 000 isoforms and (**D**) curve-based evaluation

We first compared predicted isoforms with the reference set using the number of matched transcripts to assess sensitivity. Transcripts validated against the reference set were counted. Figure 5A and B shows the number of valid transcripts and the precision, respectively, for MCF-7 cell line data. IntAPT identifies more valid transcripts with comparable precision. With the support from multiple samples, StringTie and Cuffmerge performed best in sensitivity among the current methods. However, IntAPT identified 1738 (15 584 versus 13 846) and 1757 (15 584 versus 13 827) more isoforms than Cuffmerge and ISP, respectively. With respect to precision, IntAPT had an increase of 8.51% and 12.10% over ISP and Cuffmerge, respectively. Noisy reads led FlipFlop to predict erroneously a large number of isoforms. TACO and StringTie tended to predict fewer isoforms with very high precision. Nevertheless, given the different schemes used, it is inappropriate to compare TACO and StringTie directly with the other tools based on the number of valid transcripts and precision.

To conduct a more comprehensive comparison, we calculated the precision of the top 10 000 predictions for all assemblers (Fig. 5C). IntAPT achieved a precision about 0.68, which is 11.24% higher than StringTie, 24.17% higher than Cuffmerge, 46.44% higher than Cufflinks-pool and 47.08% higher than TACO. Supplementary Figures S2-7 and S2-8 show the performance of each assembler on the H1-hESC and HepG2 data, respectively. Results confirmed that IntAPT consistently identified more valid transcripts, and with higher precision, than existing methods. Detailed results are summarized in Supplementary Table S2-3. Among all competing methods, only Cufflinks-pool and FlipFlop constructed the splicing graph on pooled data from multiple replicates. However, the large amount of noise accumulated could produce erroneous graphs, largely degrading the performance of both Cufflinks-pool and FlipFlop.

We further evaluated tool performance using a curve-based method introduced in Maretty *et al.* (2014) to adjust for possible transcript abundance bias between assemblers. We set different thresholds for the estimated isoform abundance from high to low to generate a curve between the number of predicted transcripts and the number of valid transcripts. Curve slope and height estimate precision and sensitivity, respectively. As Cuffmerge and StringTie only reported the isoform structure, we used Cufflinks to estimate the isoform abundance on the pooled data given the predicted isoform structure. Figure 5D and Supplementary Figures S2-7D and S2-8D show the curves of all assemblers on MCF-7, H1-hESC and HepG2 cell line data, respectively. As seen from the figures, the curve for

IntAPT increased faster than the other curves and finally reached the highest point among all assemblers. This curve analysis also showed that IntAPT had a higher precision when generating different number of isoforms by drawing a vertical line crossing the line plot for each method. Furthermore, drawing a horizontal line can also show that IntAPT needed less number of predicted isoforms given the same sensitivity level.

To demonstrate the effectiveness of the transcript assemblers on real tumor tissue studies, we further conducted a case study using TCGA Glioblastoma Multiforme (GBM) RNA-seq data. Analysis of high-throughput expression profiles by Verhaak *et al.* (2010) has discovered four molecular subtypes of GBM, namely Proneural (PN), Neural (N), Classical (CL) and Mesenchymal (M). These datasets are described in Supplementary Section S1.14. Among existing methods, only Cuffmerge and StringTie succeeded without runtime errors; the other methods could not complete the analysis of the data (error messages are described in Supplementary Section S1.15). Supplementary Table S2-5 lists the total number of isoforms identified from each subtype.

Because the true isoforms are unknown, we analyzed the data based on the subtype signatures assigned by Verhaak *et al.* who reported 840 genes with distinct expression profiles across the four subtypes. Gene ontology analysis identified 554 biologically meaningful signature genes that are highly expressed. For each subtype, we compared the genes corresponding to predicted isoforms with the signature genes. Supplementary Table S2-6 shows the number of signature genes identified by Verhaak *et al.*, IntAPT, Cuffmerge and StringTie. All assemblers had similar performances and identified >88% of the signature genes reported by Verhaak *et al.* We then compared the identified isoforms from the signature genes with the ENSEMBL annotation on human GRCh38 genome assembly (version 87) (Yates *et al.*, 2015) using Cuffcompare (the details of Cuffcompare codes of isoforms relationships are described in Supplementary Section S1.16). The isoforms are categorized into three groups: (i) intron-chain match ('=' category), (ii) novel splicing ('j' category) and (iii) false predictions ('other' category). Supplementary Figures S2-9 and S2-10 show the categories of isoforms identified by IntAPT, Cuffmerge and StringTie, which all identified a similar number of isoforms validated against the assigned annotations for different subtypes. However, IntAPT and StringTie had a much higher precision than Cuffmerge (as shown in Supplementary Fig. S2-11). Supplementary Figure S2-11 also shows that the low precision of Cuffmerge mainly came from a large number of isoforms in 'j' and 'other' categories, which, consistent with our simulation study, are likely due to the accumulation of individual false positive junctions using Cuffmerge. We also studied the genes with potential novel splicing isoforms identified by IntAPT using the DAVID Functional Annotation Tool (Huang *et al.*, 2009). Table 1 shows the enriched signaling pathways and functions for each subtype, which are closely related to the development of GBM.

In addition to evaluations at the gene level, we validated the IntAPT-predicted isoforms in the '=' category using an independent RT-qPCR dataset. Pal *et al.* (2014) performed RT-qPCR analysis on an independent cohort of GBM samples from the University of Pennsylvania tissue bank. The RT-qPCR data include the measured expression of 164 isoforms (including 38 signature genes) across 226 samples. We found 11 isoforms of the 38 signature genes were included in the '=' category identified by IntAPT; these 11 isoforms (see Supplementary Table S2-8) were successfully validated by RT-qPCR as the signature genes for GBM subtypes. For example, overexpression of the Ras signaling pathway has been widely observed in GBM, and is a potential target for glioma treatment (Mao *et al.*, 2012). Activation of NFkB signaling pathway is also related to the initiation of cell proliferation in GBM. Furthermore, we compared the novel isoforms of each subtype with the identified isoform sets of the other subtypes. Supplementary Figure S2-12 shows the number of novel isoforms that only existed in one subtype (Supplementary Fig. S2-13).

Table 1. Functional annotation of genes with potential novel splicing isoforms

Subtype	Functions category	Functions	*P*-value
CL	KEGG pathway	Glioma	3.00E-4
	GOTERM	Regulation of transcription	1.40E-3
	KEGG pathway	Ras signaling pathway	5.20E-3
M	GOTERM	Regulation of cell proliferation	4.80E-3
	KEGG pathway	NFkB signaling pathway	1.20E-2
N	GOTERM	Extracellular exosome	1.90E-3
	GOTERM	Oxidative phosphorylation	3.30E-2
PN	GOTERM	Cell junction	2.80E-3
	GOTERM	Cell adhesion	1.10E-2
	GOTERM	Positive regulation of GTPase activity	3.70E-2

4 Discussion

We have developed a probabilistic approach to identify phenotype-specific isoforms from multiple RNA-seq profiles. A key aspect of our approach is two-layer Bayesian modeling of phenotype-level isoform presence and abundance across multiple samples. At the group layer, we model each isoform's presence as being expressed or unexpressed. Gibbs sampling iteratively estimates the presence and abundance of isoforms. Compared with previously developed methods, this sampling framework allows us to quantify isoform presence and abundance concurrently. Based on sampling frequencies, IntAPT reports, for each isoform, a confidence measure of its presence and abundance. Another advantage of Gibbs sampling is improved identification of low abundance isoforms, which is also address by another Bayesian assembler (Aguiar *et al.*, 2018). IntAPT demonstrated improved sensitivity for lowly expressed isoforms in both our simulation and real data studies.

IntAPT can be applied to multiple samples sharing a consistent phenotype, with the definition of phenotype varying according to the experimental design. For example, as presented here, it can identify either subtype-specific isoforms or, by taking all samples from the TCGA GBM dataset, more generic GBM-specific isoforms. To demonstrate IntAPT's ability to analyze datasets consisting of a larger number of samples, we conducted a realistic simulation study on chr1 of 51 samples. The expression profiles of these samples are from CCLE breast cancer cell lines. In the realistic simulation shown in Supplementary Figure S2-14, IntAPT achieved much higher precision due to its robustness to noise, which is consistent with our findings on a small-scale realistic simulation (Supplementary Fig. S2-6). We have demonstrated that our proposed model has improved performance over datasets with various sample size by analyzing multiple samples simultaneously, but we would not suggest using our package to analyze a huge number of samples because this will increase the computational burden on the splicing graph construction. For real applications, a dataset with <100 samples will be ideal for our package, which should be enough for most biological studies and databases with detailed phenotypes.

Funding

This work was supported in part by the National Institutes of Health [CA149653 to J.X., CA164384 to L.H.-C., CA149147 & CA184902 to R.C., CA148826 & CA187512 to T.-L.W. and GM125878 to A.F.N.].

Conflict of Interest: The authors declare that there is no conflict of interest.

References

Aguiar,D. *et al.* (2018) Bayesian nonparametric discovery of isoforms and individual specific quantification. *Nat. Commun.*, **9**, 1681.

Bernard,E. *et al.* (2015) A convex formulation for joint RNA isoform detection and quantification from multiple RNA-seq samples. *BMC Bioinformatics*, **16**, 262.

Cancer Genome Atlas Network. (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.
Conesa,A. *et al.* (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol.*, **17**, 13.
Damien,P. and Walker,S.G. (2001) Sampling truncated normal, beta, and gamma densities. *J. Comput. Graph. Stat.*, **10**, 206–215.
Djebali,S. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.
Dobin,A. *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
Frazee,A.C. *et al.* (2015) Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics*, **31**, 2778–2784.
Gonzalez-Garay,M.L. (2016) Introduction to isoform sequencing using pacific biosciences technology (Iso-Seq). In: *Transcriptomics and Gene Regulation*. Springer, Dordrecht, pp. 141–160.
Griebel,T. *et al.* (2012) Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Res.*, **40**, 10073–10083.
Hah,N. *et al.* (2011) A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell*, **145**, 622–634.
Holzer,M. and Marz,M. (2019) De novo transcriptome assembly: a comprehensive cross-species comparison of short-read RNA-Seq assemblers. *Gigascience*, **8**, giz039.
Huang,D.W. *et al.* (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
Ishwaran,H. and Rao,J.S. (2005) Spike and slab variable selection: frequentist and Bayesian strategies. *Ann. Stat.*, **33**, 730–773.
Kim,D. *et al.* (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36.
Kimmig,A. *et al.* (2012) A short introduction to probabilistic soft logic. In: *Proceedings of the NIPS Workshop on Probabilistic Programming: Foundations and Applications, Lake Tahoe, Nevada, USA*. pp. 1–4.
Li,J.J. *et al.* (2011) Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. *Proc. Natl. Acad. Sci. USA*, **108**, 19867–19872.
Li,W. and Jiang,T. (2012) Transcriptome assembly and isoform expression level estimation from biased RNA-Seq reads. *Bioinformatics*, **28**, 2914–2921.
Mao,H. *et al.* (2012) Deregulated signaling pathways in glioblastoma multiforme: molecular mechanisms and therapeutic targets. *Cancer Investig.*, **30**, 48–56.
Maretty,L. *et al.* (2014) Bayesian transcriptome assembly. *Genome Biol.*, **15**, 501.
Martin,J.A. and Wang,Z. (2011) Next-generation transcriptome assembly. *Nat. Rev. Genet.*, **12**, 671–682.
McIntyre,L.M. *et al.* (2011) RNA-seq: technical variability and sampling. *BMC Genomics*, **12**, 1.
Mitchell,T.J. and Beauchamp,J.J. (1988) Bayesian variable selection in linear regression. *J. Am. Stat. Assoc.*, **83**, 1023–1032.
Niknafs,Y.S. *et al.* (2017) TACO produces robust multisample transcriptome assemblies from RNA-seq. *Nat. Methods*, **14**, 68–70.
Pal,S. *et al.* (2014) Isoform-level gene signature improves prognostic stratification and accurately classifies glioblastoma subtypes. *Nucleic Acids Res.*, **42**, e64.
Pertea,M. *et al.* (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, **33**, 290–295.
Pruitt,K.D. *et al.* (2014) RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.*, **42**, D756–D763.
Rhoads,A. and Au,K.F. (2015) PacBio sequencing and its applications. *Genom. Proteom. Bioinf.*, **13**, 278–289.
Robinson,M.D. and Smyth,G.K. (2007) Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, **23**, 2881–2887.
Robinson,M.D. *et al.* (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
Rosenbloom,K.R. *et al.* (2015) The UCSC genome browser database: 2015 update. *Nucleic Acids Res.*, **43**, D670–D681.
Shi,X. *et al.* (2018) SparseIso: a novel Bayesian approach to identify alternatively spliced isoforms from RNA-seq data. *Bioinformatics*, **34**, 56–63.
Shi,Y. (2017) Mechanistic insights into precursor messenger RNA splicing by the spliceosome. *Nat. Rev. Mol. Cell Biol.*, **18**, 655–670.
Tasnim,M. *et al.* (2015) Accurate inference of isoforms from multiple sample RNA-Seq data. *BMC Genomics*, **16**, S15.
Trapnell,C. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.
Verhaak,R.G. *et al.* (2010) Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, **17**, 98–110.
Wang,E.T. *et al.* (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.
Yates,A. *et al.* (2015) Ensembl 2016. *Nucleic Acids Res.*, **44**, D710–D716.